2P4-2

逆強化学習における制約条件の緩和法

Relaxation of constraint conditions in inverse reinforcement learning

北里 勇樹 *1　荒井 幸代 *1
Yuki Kitazato　Sachiyo Arai

*1千葉大学大学院工学研究科都市環境システムコース
Graduate School of Engineering, Chiba University, Division of Urban Environment Systems

Inverse Reinforcement Learning (IRL) is a promising framework for estimating a reward function under the given optimal policy. An original idea of IRL is proposed by Russell et.al. where the constraint conditions are calculated by comparison of the difference of maximum Q-value and other Q-value of a state over the whole states. In a large-scale environment, it becomes difficult to solve by increasing the number of constraints due to the increase in the number of states. In this paper we propose a relaxation method of constraint conditions in inverse reinforcement learning. In the proposed method, important constraint conditions extracts from original constraint conditions and reconstructs the optimization problem using them. Then, we show the effectiveness of our method via maze problem.Through computer experiment, we found that the important constraint conditions exists the state where there are multiple optimum behaviors close to the start state.

1. はじめに

代表的な逆強化学習の一つである Ng の逆強化学習 [1] は制約付き最適化問題として定式化される．最適方策の獲得を保証するための制約条件は価値関数に基づいて計算されるが，大規模な環境においては，状態数に比例した制約条件数の増加により，問題の解決が困難となる．そこで本稿では制約条件の中から特に重要な制約条件を抽出し，制約条件を緩和するための手法を提案する．

2. 準備

本稿の理解に必要なマルコフ決定過程，強化学習の基礎理論に加え，本稿で注目する逆強化学習である Ng[1] の手法の説明と記号の定義を行う．

2.1 マルコフ決定過程（MDP）

マルコフ決定過程は状態遷移にマルコフ性を持つ動的最適化のための数学モデルである．マルコフ性とは次状態への遷移が直前の状態と行動のみに依存するという性質である．有限マルコフ決定過程は $\langle S, A, P^a_{ss'}, \gamma, R\rangle$ からなる．S は有限状態集合，A は行動集合，$P^a_{ss'}$ は状態 $s \in S$ で行動 $a \in A$ ととったとき次状態 $s' \in S$ に遷移する確率，γ は割引率，R は報酬関数を表す．

2.2 強化学習

強化学習 [2] は，未知の環境において最適な制御則を試行錯誤的に獲得する手法である．意思決定主体であるエージェントには，状態入力に対する正しい出力を明示した教師信号が存在せず，報酬と呼ばれるスカラーの情報のみが与えられる．エージェントはこの報酬の期待総和を最大化することを目的とし，学習を行う．環境モデルを $\langle S, A, R, \pi\rangle$ と定義する．方策 π は状態 s から可能な行動 a を選択する確率である．エージェントは時刻 t において状態 $s_t \in S$ を観測し，自身の方策 π_t に基づいて行動 $a_t \in A$ を選択する．その後，時刻 $(t+1)$ では s_t，a_t によって確率的に次状態 s_{t+1} に遷移し，報酬 r_t を得る．獲得した報酬から価値関数 $V(s)$ または行動価値関数 $Q(s,a)$ を生成し，その値を用いて方策 π の評価と改善を行う．価値関数 $V(s)$，行動価値関数 $Q(s,a)$ は方策 π が与えられたとき，それぞれ式 (1)，式 (2) を満たす．

$$V^\pi(s) = R(s) + \gamma \sum_{s'} P_{s\pi(s)}(s') V^\pi(s') \tag{1}$$

$$Q^\pi(s,a) = R(s) + \gamma \sum_{s'} P_{sa}(s') V^\pi(s') \tag{2}$$

2.3 Ng の逆強化学習

Ng らの逆強化学習 [1] は，MDP から R を除く $\langle S, A, P^a_{ss'}, \gamma\rangle$ と各状態 s における最適行動 a_1 を所与とし，報酬関数 R を推定する．最適行動 a_1 は式 (3) となる．

$$a_1 \equiv \pi(s) \in \arg\max_{a \in A} R(s) + \sum_{s'} P_{sa}(s') V^\pi(s') \quad \forall s \in S \tag{3}$$

各状態において最適行動の Q 値が最大となることから，式 (4) が導かれる．

$$\sum_{s'} P_{sa_1}(s') V^\pi(s') \geq \sum_{s'} P_{sa}(s') V^\pi(s') \quad \forall s \in S, a \in A \tag{4}$$

ここで，$P_{ia}(j)$ を (i,j) 成分とする $M \times M$ 行列を状態遷移行列 $\boldsymbol{P}_a$，状態価値関数ベクトル $\boldsymbol{V}^\pi = \{V^\pi(s_i)\}_{i=0}^M$，報酬関数ベクトル $\boldsymbol{R} = \{R(s_i)\}_{i=0}^M$ を定義すれば，式 (4)，式 (1) はそれぞれ式 (5)，式 (6) と書き直せる．

$$\boldsymbol{P}_{a_1} \boldsymbol{V}^\pi \geq \boldsymbol{P}_a \boldsymbol{V}^\pi \tag{5}$$

$$\boldsymbol{V}^\pi = \boldsymbol{R} + \gamma \boldsymbol{P}_{a_1} \boldsymbol{V}^\pi \tag{6}$$

連絡先: 北里勇樹，千葉大学大学院工学研究科，千葉市稲毛区弥生町 1-33，043-290-3316

式 (6) を $\boldsymbol{V}^{\pi}$ について解き，式 (5) に代入した後に整理すると，式 (7) が導かれる．

$$(\boldsymbol{P}_{a_1} - \boldsymbol{P}_a)(\boldsymbol{I} - \gamma \boldsymbol{P}_{a_1})^{-1}\boldsymbol{R} \geq 0 \tag{7}$$

したがって，式 (7) は最適行動の Q 値が最大となることを保証するため条件である．Ng らは式 (7) を制約条件，式 (7) を各状態ごとに最大化し，報酬の総量を制御するためのペナルティ係数を加えた目的関数を設定し，逆強化学習を線形計画問題として定式化した．

$$\text{maximize} : \sum_{i=1}^{N} \min_{a \in a_2, \ldots, a_k} \{(\boldsymbol{P}_{a1}(i) - \boldsymbol{P}_a(i)) (\boldsymbol{I} - \gamma \boldsymbol{P}_{a1})^{-1}\boldsymbol{R}\} - \lambda ||\boldsymbol{R}||_1 \tag{8}$$

$$\text{subject to} : (\boldsymbol{P}_{a_1} - \boldsymbol{P}_a)(\boldsymbol{I} - \gamma \boldsymbol{P}_{a_1})^{-1}\boldsymbol{R} \geq 0 \quad \forall a \in A \backslash a_1$$

3. 制約緩和問題

大規模な環境では状態数に比例した制約条件数の増加により逆強化学習の適用が困難になる．Ng らの定式化において，制約条件に類似した目的関数を用いることから，多くの制約条件は不要となることが考えられる．そこで制約緩和問題は，元の制約条件による定式化の解と同等の解を得ることができる定式化のうち，制約条件数を最小とする問題とする．

4. 提案手法

本提案では，制約条件の中から重要なものを抽出し，それらを組み合わせることにより，新たな最適化問題を定式化する．ここで，重要な制約条件とは，解に影響を与えるものとする．具体的な手法は，次の 4 ステップからなる．

1. 制約条件を除外する個数 n の決定
2. すべての制約条件の中から n 個を除外したすべての組み合わせに対して式 (8) を解く
3. 除外した制約条件の組み合わせのうち同じ報酬関数が得られるパターンを抽出
4. 各パターンの中で出現頻度の高い制約条件を抽出

4 ステップを通して抽出した重要な制約条件を新たな制約条件として式 (8) を再定式化することにより，制約条件の緩和を行う．

5. 計算機実験

強化学習のベンチマークとして用いられる迷路問題を対象に実験を行う．実験環境を図 1 に示す．図中の矢印は所与とした最適行動である．

スタートが左上の座標 (0,0)，ゴールが右下の座標 (4,4) として，スタートからゴールまでの最短経路を発見する問題である．逆強化学習のパラメータは，報酬の値域 $-1 < R < 1$，割引率 $\gamma = 0.9$，ペナルティ係数 $\lambda = 0$ とした．

緩和する制約条件の数 n を 1 から 4 まで変化させたときの獲得したパターン数を表 1 に示す．表 1 より，緩和数が増加

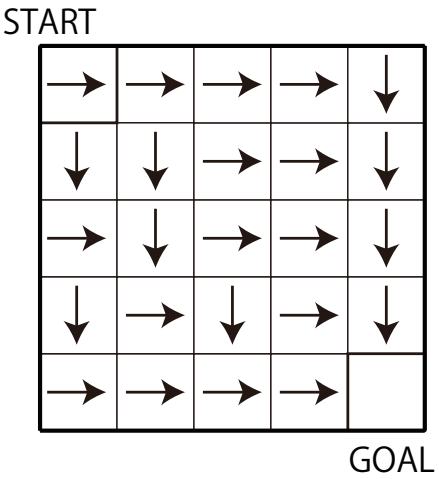


図 1: 実験環境

表 1: 制約条件の緩和数による獲得パターンの変化

緩和数	パターン数
1	2
2	5
3	8
4	8

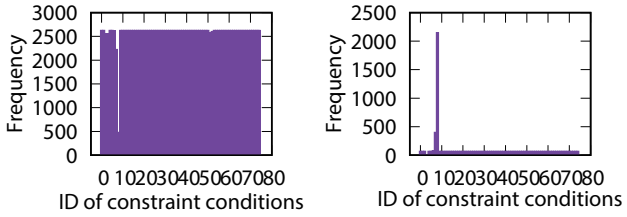
するごとに得られるパターン数も増加する．緩和数が 3 と 4 では得られるパターン数に違いが生じなかったため，以後，緩和する制約条件の数 $n = 3$ を採用した結果について述べる．

制約条件を緩和した際の各制約条件の出現頻度をパターンごとに図 2 に示す．横軸が制約条件の ID，縦軸が出現頻度である。なお各制約条件は便宜上，最適行動と二番目にいい行動との比較を座標 (0,0)，(0,1)，… の順に ID を 0,1, … としており，二番目にいい行動との比較が終わったのち三番目，四番目の行動の比較という順になっている．図 2 から各パターンはいくつかの制約条件の影響によって出現することがわかる．それぞれのパターンの出現頻度の高い制約条件を表 2 に示す．パターン 0 については出現した制約条件が多数存在するため，表中には示していない．

表 2 から，重要な制約条件は 2, 3, 7, 8, 51, 52 の 6 個であることがわかる．この制約条件を用いて最適化問題を最適式化

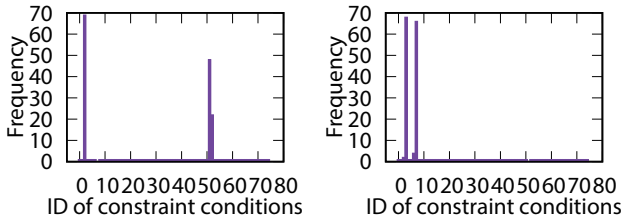
表 2: 各パターンにおける重要な制約条件

パターン	制約条件の ID
0	-
1	7, 8
2	2, 51, 52
3	3, 7
4	3, 8
5	2, 7, 51
6	3, 7, 51
7	3, 8, 51



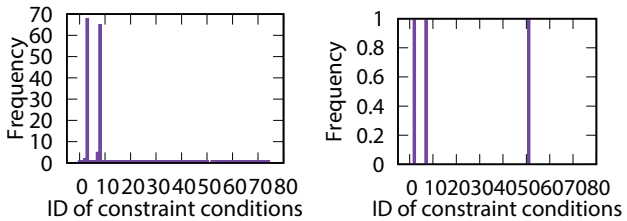
(a) pattern 0

(b) pattern 1



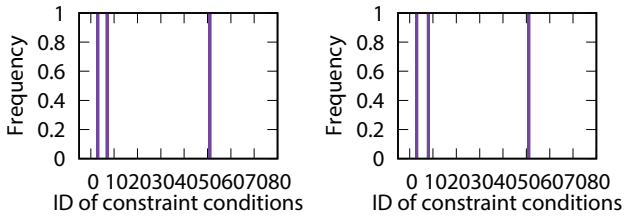
(c) pattern 2

(d) pattern 3



(e) pattern 4

(f) pattern 5



(g) pattern 6

(h) pattern 7

図 2: 各パターンにおける制約条件の出現頻度

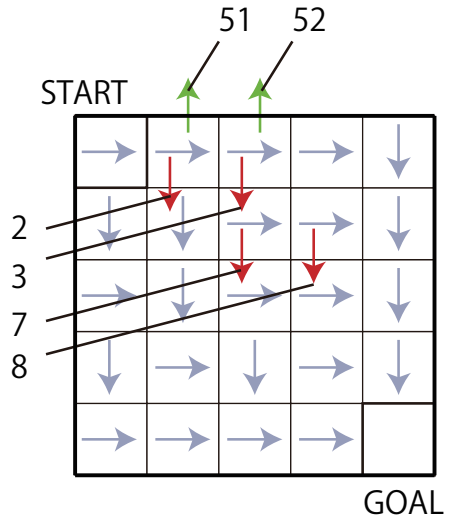


図 3: 重要な制約条件と迷路問題の対応

し解いた結果，すべての制約条件がある場合と同じ解が得られた．

6. 考察

重要な制約条件として抽出された 2, 3, 7, 8, 51, 52 について詳しく見る．一つの制約条件はある状態において，最適行動とある行動の Q 値を比較するものである．重要な制約条件として抽出されたものがどの状態行動対との比較を表しているかを図 3 に示す．図 3 から以下の二つの条件が読み取れる．

- スタート付近の状態
- 最適行動が二つ以上の状態

通常目標状態であるゴールに対して大きな報酬が与えられるため，ゴールの報酬から割り引かれた報酬が Q 値となり，各行動の Q 値がスタート地点付近では相対的に小さくなる．Ng らの定式化における目的関数は最適行動とそのほかの行動の Q 値の差が生じやすい状態に注目するため，これらの状態の制約条件を遵守する必要性が生じたと考えられる．最適行動が二つ以上の状態では，複数の最適行動の Q 値の差が生じにくく，逆転しやすい状態であるといえるため，制約条件の影響が大きくなったと考えらえる．

7. まとめ

本稿では，Ng らが線形計画問題として提案した逆強化学習の制約条件に注目し，制約条件を緩和することにより問題の複雑さを減少するための手法を提案した．提案手法では，制約条件の組み合わせを網羅的に探索し，重要な制約条件を抽出した．重要な制約条件の持つ性質として，スタート付近の状態かつ最適行動が二つ以上ある状態の制約条件が重要であるという知見が得られた．本稿では一つの問題だけを取り扱ったため，今後は様々な問題に対して提案手法の有用性を検証する必要がある．また，本手法は制約を見つける段階で最適化問題を複数回解く必要があるため，得られた知見を利用して，最適化を解く回数を少なくする必要がある．

参考文献

[1] Andrew Y. Ng，Stuart Russell: Algorithms for Inverse Reinforcement Learning，In Proceedings of the Seventeenth International Conference on Machine Learning，pp.663-670，(2000)

[2] Richard S. Sutton, Andrew G. Barto: Reinforcement Learning: An Introduction, 三上貞芳，皆川 雅章訳: ”強化学習”, 森北出版， pp.142-170, (2000)